KMANPUB | AI and Tech in Behavioral and Social Sciences Vol. 4 No. 4 (2026): 1-7

Article history:
Received 05 April 2026
Revised 01 July 2026
Accepted 07 July 2026
Initial Publication 12 July 2026
Final Publication 01 October 2026

# Supervised Classification of Civil Unrest-Related Posts in Twitter/X Data: Evidence from the CUT Dataset

Pouya. Sohofi[1], Amir. Hossein Kabiri Nameghi[1], Hassan. Naderi[2*]

1. Department of Computer Engineering, Iran University of Science and Technology, Tehran, Iran
2. Faculty Member, Department of Software Engineering, Iran University of Science and Technology, Tehran, Iran

*** Corresponding author email address:** naderi@iust.ac.ir

A r t i c l e  I n f o

**Article type:**
*Original Research*

**How to cite this article:**
Sohofi, P., Hossein Kabiri Nameghi, A., & Naderi, H. (2026). Supervised Classification of Civil Unrest-Related Posts in Twitter/X Data: Evidence from the CUT Dataset. *AI and Tech in Behavioral and Social Sciences, 4*(4), 1-7. https://doi.org/10.61838/kman.aitech.5791



A B S T R A C T

Social media platforms provide rapid public reports during demonstrations, crises, and civil unrest, but the volume and noise of user-generated content make manual monitoring impractical. This retrospective text-classification study evaluated supervised machine learning models for identifying incident-related civil unrest posts using the Civil Unrest on Twitter (CUT) dataset. The dataset consisted of 4,381 manually annotated English-language Twitter posts collected from 42 countries between 2014 and 2019. The analysis used keyword-based collection, language filtering, crowdsourced annotation, unigram bag-of-words representation, class-balancing procedures reported for the dataset, and supervised classification. Five algorithms were evaluated on the selected balanced dataset: Naive Bayes, Support Vector Machine, Logistic Regression, Gradient Boosted Decision Trees, and Convolutional Neural Network, each with and without hashtag features. In the original distribution, 690 posts were incident-related and 3,691 were non-incident-related. Dataset 2, containing 6,978 observations with 27% incident-related and 73% non-incident-related posts, was selected because it produced stronger minority-class performance. Incident-related F1-scores ranged from 0.845 to 0.915, and AUC values ranged from 0.958 to 0.977. SVM Model 1, trained with hashtag features, achieved the highest incident-related F1-score (0.915). The findings suggest that supervised classification may provide a useful filtering layer for analyst-supported civil unrest monitoring. However, the framework was evaluated on historical batch data rather than in a prospective streaming setting; therefore, real-time validation, transparent governance, multilingual testing, and stronger ethical safeguards are required before operational deployment.

***Keywords:*** *civil unrest; Twitter/X; protest demonstrations; incident classification; machine learning; social media analytics; Support Vector Machine*

## 1. Introduction

Civil unrest and protest demonstrations increasingly unfold in parallel with intense digital activity. Participants, witnesses, journalists, public authorities, and ordinary observers publish short messages that describe events, express opinions, circulate links, and react to rapidly changing situations. Twitter, now X, is especially relevant for this purpose because its public microblogging format supports rapid diffusion through hashtags, mentions, and reposting practices. Previous research has shown that social media can influence political participation and collective action, while also providing data that can be mined for event detection and crisis response (Boulianne,

2015; Enikolopov et al., 2020; Imran et al., 2015; Van Laer & Van Aelst, 2010). For public safety and emergency management, the value of such information depends on whether meaningful signals can be separated from the much larger stream of unrelated, ambiguous, or purely opinion-based content.

Early warning of civil unrest-related incidents is not identical to general event detection. A monitoring system must identify posts that refer to a relevant incident during or around a protest, not simply detect increased discussion volume. This distinction is important because civil unrest conversations often include background dissatisfaction, political commentary, news sharing, and symbolic language. Only a smaller subset of messages directly refers to specific incidents such as violence, confrontation, disruption, repression, or emerging risk. Automated detection therefore requires supervised methods trained on annotated examples, consistent with broader work on Twitter event detection and crisis informatics (Atefeh & Khreich, 2015; Imran et al., 2015; Marcus et al., 2011).

The present study examines whether supervised text-classification models can distinguish incident-related from non-incident-related civil unrest posts in the CUT dataset. The study has three aims: first, to describe the dataset and classification pipeline; second, to report the distribution of incident-related posts and the class-balancing strategy; and third, to compare the performance of machine learning models used for incident-related classification. Because the analysis is based on historical Twitter data, the term early warning is used cautiously to refer to a potential decision-support filtering layer rather than a fully validated real-time warning system.

## 2. Methods and Materials

### *2.1. Study Design and Data Source*

This study used a retrospective quantitative text-classification design based on the Civil Unrest on Twitter (CUT) dataset. The CUT dataset is a manually annotated corpus of 4,381 English-language Twitter posts related to civil unrest, collected from 42 countries between 2014 and 2019. The analytical objective was binary classification of posts as incident-related or non-incident-related. In this manuscript, incident-related posts refer to messages that directly or non-specifically refer to civil unrest events, whereas broader expressions of dissatisfaction are treated as a separate annotation dimension rather than as automatic evidence of an incident.

All data were obtained from the CUT dataset. The original collection used the Twitter Streaming API with geolocation filters targeting African, Middle Eastern, and Southeast Asian regions. Messages were retrieved using 709 English-language keywords associated with unrest and political conflict, including terms such as "unemployment," "police," and "extremist." Only English-language messages were retained. Language filtering was performed using an external language identification tool, retweets were excluded, and all retained messages contained geolocation metadata due to the collection method. The initial set contained 4,415 tweets; after removal of 34 non-English items, 4,381 annotated messages remained.

### *2.2. Preprocessing, Annotation, and Feature Representation*

Preprocessing followed the protocol used during creation of the CUT dataset. Text was normalized, lowercased, and tokenized. Keyword-based filtering had already been performed during collection; therefore, no additional filtering or deduplication was applied at the modeling stage. For machine learning, posts were represented using unigram bag-of-words features, a standard representation for text classification and information retrieval tasks (HaCohen-Kerner et al., 2020; Manning et al., 2008; Uysal & Gunal, 2014). Two feature configurations were evaluated: Model 1 retained hashtag content as part of the text representation, whereas Model 2 excluded hashtag content.

Annotation was performed through a crowdsourced process in which each message was independently labeled by three workers. The annotation schema captured several dimensions, including whether a message referred to a protest, strike, or riot; whether it expressed civil or political dissatisfaction; the temporal relation of the message to an event; the stance of the user; the presence of event-related topics; intention to participate; and event-specific hashtags. Quality control included pre-annotated gold-standard items, expert adjudication of conflicting labels, and removal of low-quality annotators. Because incident-related tweets were initially underrepresented, a Random Forest model trained on early annotations was used to prioritize more relevant messages in subsequent annotation rounds. The binary civil unrest relevance label was used as the ground truth for classification.


AITBSS
AI and Tech in Behavioral and Social Sciences
E-ISSN: 3041-9433

### 2.3. *Class Balancing and Model Evaluation*

Exploratory analysis described message distributions across annotation categories, countries, and time. To address class imbalance, two class-balanced datasets were prepared as reported for the CUT analysis. Dataset 1 contained 6,499 observations, with 78% non-incident-related and 22% incident-related posts. Dataset 2 contained 6,978 observations, with 73% non-incident-related and 27% incident-related posts. Dataset 2 was selected for model training because it produced superior F-measure performance for the incident-related class. The available dataset documentation reports the final class distributions but does not fully specify the balancing mechanism; therefore, the reported results are interpreted strictly within the validation setting described for the dataset.

The evaluated classifiers were Naive Bayes, Support Vector Machine, Logistic Regression, Gradient Boosted Decision Trees, and Convolutional Neural Network. Performance was assessed using F-measure, precision, recall, AUC, and accuracy. The available performance table reports training and validation F1-scores but does not provide a fully documented external test set, cross-validation scheme, or hyperparameter-search protocol. Therefore, model performance is interpreted as historical validation performance rather than evidence of operational real-time deployment.

### 2.4. *Analyst-Supported Warning Procedure*

The proposed system architecture used the classifier as a filtering component within an analyst-supported warning pipeline. The procedure should be understood as a conceptual decision-support workflow. It was not prospectively validated on live streaming data in the present study; therefore, future operational evaluation should estimate detection delay, false-alarm rate, adaptive-threshold calibration, and performance across languages and regions.

**Algorithm 1**

*Conceptual analyst-supported warning procedure*

| Step | Operation |
|---|---|
| 1 | Collect an incoming Twitter/X post and add it to the processing queue. |
| 2 | Clean and normalize the post text. |
| 3 | Assign the cleaned post to the current time window. |
| 4 | Detect an event if the current time window exceeds the adaptive threshold and is larger than the previous window. |
| 5 | Classify posts in the detected event window. |
| 6 | If a post is classified as incident-related, mark the event as an incident and route it for analyst-supported warning review. |

## 3. Findings and Results

The original CUT dataset contained 4,381 annotated English-language posts collected from 42 countries between 2014 and 2019. Among these posts, 539 were labeled as referring to a specific civil unrest event and 151 referred to unrest in a non-specific manner. In total, 690 posts, representing 16% of the corpus, were classified as incident-related. The remaining 3,691 posts, representing 84% of the corpus, were classified as non-incident-related (Table 1; Figure 1). A broader annotation dimension showed that 1,951 posts expressed civil or political dissatisfaction, whereas 2,446 did not. These results indicate that general dissatisfaction appeared more frequently than explicit references to specific unrest events.

The initial class distribution was therefore imbalanced. To reduce the effect of imbalance on minority-class learning, two class-balanced datasets were constructed. Dataset 1 contained 6,499 observations, with 22% incident-related and 78% non-incident-related posts. Dataset 2 contained 6,978 observations, with 27% incident-related and 73% non-incident-related posts (Table 2; Figure 2). Dataset 2 was selected for subsequent model training because it yielded superior F-measure performance for the incident-related class.

Five machine learning algorithms were evaluated on Dataset 2: Naive Bayes, Support Vector Machine, Logistic Regression, Gradient Boosted Decision Trees, and Convolutional Neural Network. Each algorithm was tested in two configurations: Model 1 used tweet text with hashtag content, and Model 2 used tweet text without hashtag content. Across the evaluated models, incident-related F1-scores ranged from 0.845 to 0.915, while AUC values ranged from 0.958 to 0.977 (Table 3; Figure 3). Inclusion or exclusion of hashtag features did not substantially alter performance.

The best-performing classifier was SVM Model 1, which achieved an incident-related F1-score of 0.915,


AITBSS
AI and Tech in Behavioral and Social Sciences
E-ISSN: 3041-9433

precision of 0.947, recall of 0.885, AUC of 0.975, and accuracy of 0.955. SVM Model 2 followed closely, with an incident-related F1-score of 0.914 and accuracy of 0.954. However, several models showed large gaps between training and validation F1-scores, including SVM Model 1 (0.997 vs. 0.911) and SVM Model 2 (0.998 vs. 0.914), indicating possible overfitting. Consequently, SVM should be interpreted as the strongest model under the reported validation conditions, not as an operationally validated early-warning model.

**Table 1**

*Distribution of annotated posts in the original dataset*

| Category | Number of posts | Percentage |
|---|---|---|
| Incident-related | 690 | 16% |
| Non-incident-related | 3,691 | 84% |
| Total | 4,381 | 100% |

**Table 2**

*Class-balanced datasets used for model development*

| Dataset | Total observations | Incident-related | Non-incident-related |
|---|---|---|---|
| Dataset 1 | 6,499 | 22% | 78% |
| Dataset 2 | 6,978 | 27% | 73% |

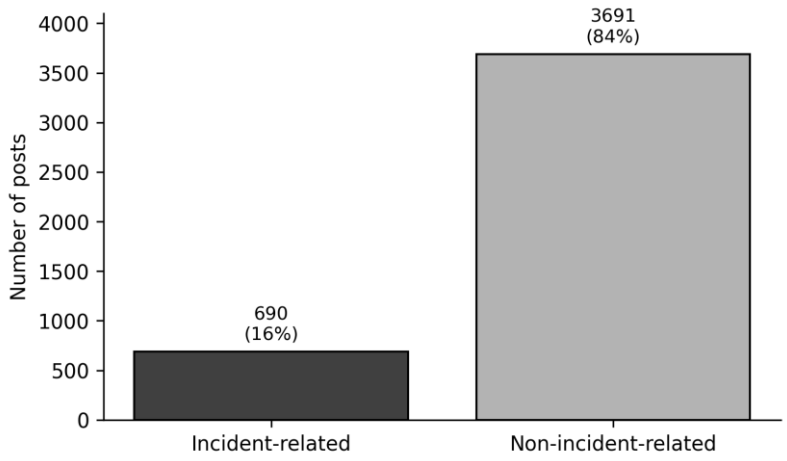
**Table 3**

*Classification performance for Dataset 2*

| Algorithm | Model | F1 train | F1 val | F1 incident | Precision incident | Recall incident | AUC | Accuracy |
|---|---|---|---|---|---|---|---|---|
| NB | 1 | 0.943 | 0.845 | 0.885 | 0.914 | 0.859 | 0.973 | 0.939 |
| NB | 2 | 0.946 | 0.850 | 0.882 | 0.904 | 0.861 | 0.974 | 0.937 |
| LR | 1 | 0.980 | 0.881 | 0.884 | 0.904 | 0.864 | 0.971 | 0.938 |
| LR | 2 | 0.995 | 0.879 | 0.884 | 0.883 | 0.885 | 0.973 | 0.936 |
| SVM | 1 | 0.997 | 0.911 | 0.915 | 0.947 | 0.885 | 0.975 | 0.955 |
| SVM | 2 | 0.998 | 0.914 | 0.914 | 0.939 | 0.890 | 0.977 | 0.954 |
| GBDT | 1 | 0.997 | 0.878 | 0.884 | 0.883 | 0.885 | 0.966 | 0.936 |
| GBDT | 2 | 0.997 | 0.878 | 0.872 | 0.880 | 0.864 | 0.964 | 0.931 |
| CNN | 1 | 0.948 | 0.895 | 0.859 | 0.893 | 0.827 | 0.958 | 0.926 |
| CNN | 2 | 0.947 | 0.910 | 0.845 | 0.890 | 0.804 | 0.958 | 0.926 |

**Figure 1**

*Distribution of incident-related and non-incident-related posts.*





AITBSS
AI and Tech in Behavioral and Social Sciences
E-ISSN: 3041-9433

**Figure 2**

*Class balance across original and balanced datasets.*

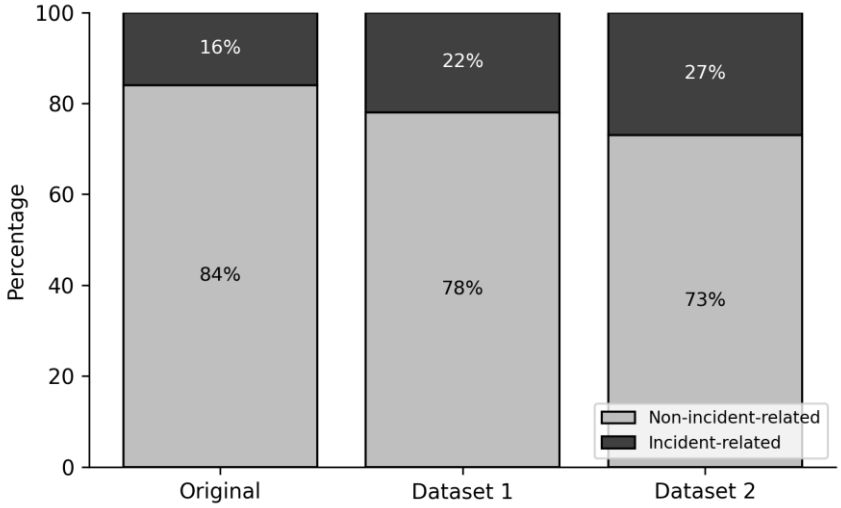
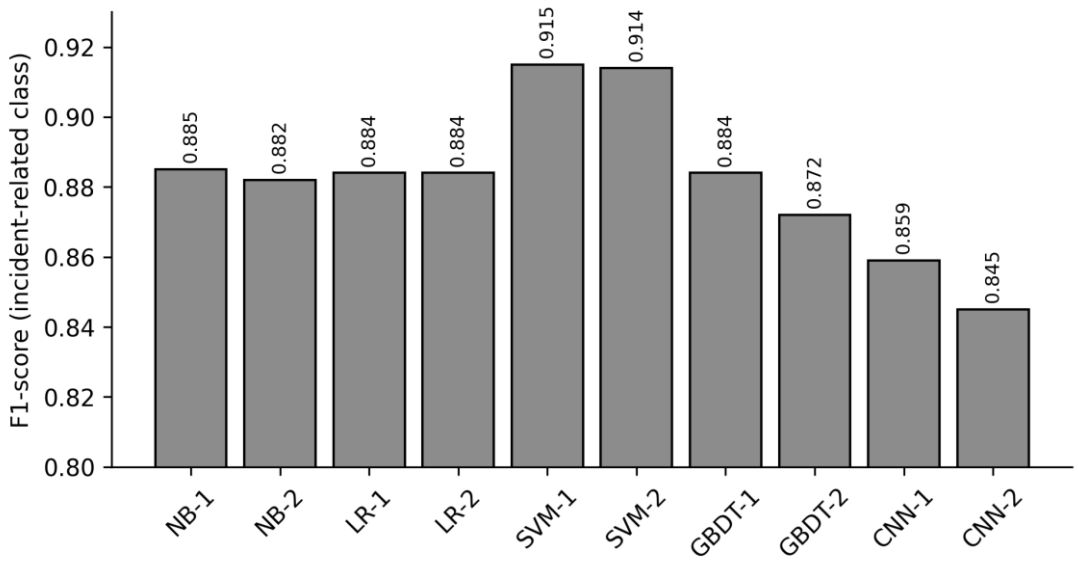



**Figure 3**

*Comparison of models by incident-related F1-score.*




## 4. Discussion

The findings support the usefulness of supervised machine learning as a filtering mechanism for civil unrest monitoring. The central practical problem is not the absence of social media data, but the imbalance between relevant and irrelevant content. Only 16% of the original posts were incident-related, showing that a monitoring system must identify a small minority of actionable messages within a much larger stream. This pattern is consistent with broader crisis informatics research, which emphasizes information overload and the need for automated filtering during emergencies (Imran et al., 2015).

The strongest reported model was SVM with hashtag features. This outcome is technically plausible because unigram bag-of-words representations produce high-dimensional sparse feature spaces, a setting in which linear SVM-based approaches have often performed well in text classification (Joachims, 1998; Manning et al., 2008). The small difference between the SVM model with hashtags and the SVM model without hashtags suggests that hashtag


AITBSS
AI and Tech in Behavioral and Social Sciences
E-ISSN: 3041-9433

content was not the dominant determinant of performance in this dataset. Instead, the classifier appears to have learned from broader textual patterns. The results also show that several conventional models achieved strong AUC values, indicating that traditional machine learning remains useful for short-text classification when manually labeled training data are available.

At the same time, the performance results should be interpreted cautiously. Large differences between training and validation F1-scores in several models indicate possible overfitting, particularly for high-dimensional sparse models. Moreover, the available documentation does not fully specify the balancing method, the complete validation protocol, or whether a fully independent test set was used. These issues limit the strength of claims that can be made about generalizability. Future work should report the exact resampling procedure, ensure that balancing is applied only within the training set, use independent test data or cross-validation, provide confidence intervals, and include precision-recall analysis because the incident-related class is the minority class.

Operationally, the proposed system should be understood as a decision-support tool rather than a fully autonomous public safety system. A classifier can reduce the burden on analysts by routing potentially relevant posts into further review, but automated warnings can be affected by sarcasm, misinformation, missing context, political bias, and uneven platform use. These risks are consistent with broader concerns about misinformation, crisis-event analysis, and automated social media monitoring (Kraft & Usbeck, 2022; Shu et al., 2017). The present evaluation used historical batch data rather than live streaming data. Before deployment, the framework would need prospective validation, calibration of warning thresholds, multilingual testing, interpretability assessment, and governance rules for responsible use.

The ethical implications of protest monitoring require particular caution. Tools designed to identify civil unrest-related posts may be useful for emergency management, but they may also create risks of surveillance, political profiling, re-identification, or disproportionate monitoring of vulnerable groups. Accordingly, any operational system should include human oversight, minimal-data principles, transparent governance, restrictions on harmful uses, and mechanisms for accountability. These safeguards are essential because protest monitoring can have direct implications for civil liberties and public trust.

## 5. Conclusion

This study evaluated supervised machine learning models for classifying incident-related civil unrest posts in the CUT dataset. The dataset contained 4,381 annotated English-language Twitter posts, of which 690 were incident-related. After class balancing, Dataset 2 was selected for model training. Across five evaluated algorithms, SVM achieved the strongest incident-related validation performance, with an F1-score of 0.915 when hashtag features were included. The results suggest that supervised classification can serve as a useful filtering layer for analyst-supported civil unrest monitoring. However, the findings should be interpreted as retrospective classification evidence rather than proof of a fully validated real-time early-warning system. Operational use requires prospective streaming validation, clearer balancing and evaluation protocols, stronger interpretability, multilingual extension, and careful ethical safeguards.

## Authors' Contributions

Pouya Sohofi: conceptualization, data analysis, writing-original draft. Amir Hossein Kabiri Nameghi: methodology, software, analysis. Hassan Naderi: supervision, validation, writing-review and editing.

## Declaration

Artificial intelligence (AI)-assisted tools were used to improve the linguistic quality, readability, and grammatical accuracy of the manuscript. The authors retained full responsibility for the study design, data collection, data analysis, interpretation of the findings, and final content. All AI-assisted outputs were reviewed, verified, and edited by the authors before submission. No AI tool was used as an author of the manuscript.

## Transparency Statement

The study is based on the Civil Unrest on Twitter (CUT) dataset and the variables and performance metrics reported for that dataset. No new dataset was generated for this manuscript. Access to the underlying data should follow the access conditions and redistribution rules of the original dataset and the Twitter/X platform.

## Acknowledgments


AITBSS
AI and Tech in Behavioral and Social Sciences
E-ISSN: 3041-9433

We would like to express our gratitude to all individuals helped us to do the project.

## Declaration of Interest

The authors report no conflict of interest.

## Funding

According to the authors, this article has no financial support.

## Ethics Considerations

This study involved secondary analysis of an existing annotated social media dataset and did not involve new direct interaction with human participants or animals. Only aggregate results are reported, and no usernames, handles, tweet identifiers, or identifiable post content are reproduced in this manuscript. Any operational use of protest-monitoring systems should follow applicable legal, institutional, and ethical requirements, including proportionality, human oversight, and safeguards against harmful use.

## References

Atefeh, F., & Khreich, W. (2015). A survey of techniques for event detection in Twitter. *Computational Intelligence*, *31*(1), 132-164. https://doi.org/10.1111/coin.12017

Boulianne, S. (2015). Social media use and participation: A meta-analysis of current research. *Information, Communication & Society*, *18*(5), 524-538. https://doi.org/10.1080/1369118X.2015.1008542

Enikolopov, R., Makarin, A., & Petrova, M. (2020). Social media and protest participation: Evidence from Russia. *Econometrica*, *88*(4), 1479-1514. https://doi.org/10.3982/ECTA14281

HaCohen-Kerner, Y., Miller, D., & Yigal, Y. (2020). The influence of preprocessing on text classification using a bag-of-words representation. *PLoS One*, *15*(5), e0232525. https://doi.org/10.1371/journal.pone.0232525

Imran, M., Castillo, C., Diaz, F., & Vieweg, S. (2015). Processing social media messages in mass emergency: A survey. *ACM Computing Surveys*, *47*(4), Article 67. https://doi.org/10.1145/2771588

Joachims, T. (1998). *Text categorization with support vector machines: Learning with many relevant features*. Springer. https://doi.org/10.1007/BFb0026683

Kraft, A., & Usbeck, R. (2022). The ethical risks of analyzing crisis events on social media with machine learning. *arXiv*. https://arxiv.org/abs/2210.03352

Manning, C. D., Raghavan, P., & Schutze, H. (2008). *Introduction to information retrieval*. Cambridge University Press. https://doi.org/10.1017/CBO9780511809071

Marcus, A., Bernstein, M. S., Badar, O., Karger, D. R., Madden, S., & Miller, R. C. (2011). TwitInfo: Aggregating and visualizing microblogs for event exploration. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, *19*(1), 22-36. https://doi.org/10.1145/3137597.3137600

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing & Management*, *50*(1), 104-112. https://doi.org/10.1016/j.ipm.2013.08.006

Van Laer, J., & Van Aelst, P. (2010). Internet and social movement action repertoires: Opportunities and limitations. *Information, Communication & Society*, *13*(8), 1146-1171. https://doi.org/10.1080/13691181003628307



AITBSS
AI and Tech in Behavioral and Social Sciences
E-ISSN: 3041-9433